Leveraging Nonnegative Matrix Factorization in Processing the Temporal Modulation Spectrum for Speech Enhancement

Syu-Siang Wang[1], Jeremy Chiaming Yang[2], Yu Tsao[2] and Jeih-weih Hung[3]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]Dept of Electrical Engineering, National Chi Nan University, Taiwan

Abstract--**This paper proposes to employ the technique of nonnegative matrix factorization (NMF) to enhance the temporal modulation components of speech signals for reducing the noisy effect. As for any arbitrary acoustic frequency, the NMF-wise bases for the temporal modulations of both the clean speech and noise are first extracted and then applied to the decomposition of the temporal modulation of the noise-corrupted speech signal. In this way the noise-free speech component can be highlighted and the updated speech signal possesses higher quality than the original counterpart. Moreover, the temporal modulations of the neighboring acoustic frequencies can be processed together to boost the computation efficiency without deteriorating the enhancement. The evaluation experiments conducted on a subset of the Aurora-2 connected digit database reveal that the proposed method significantly improves the Perceptual Evaluation of Speech Quality (PESQ) scores of the signals.**

I. INTRODUCTION

As for speech-related applications, signal enhancement in noisy situations has been constantly a major and important task because the ambient noise always worsens the corresponding performance. Generally speaking, the techniques with respect to automatic speech enhancement and noise reduction are a form of machine learning, and they can be divided into two schools: supervised learning and unsupervised learning.

The supervised SE methods adopt the pre-labeled noise data to establish models or representatives, which can be codebooks [1], hidden Markov models [2] and spectral basis matrices based on the technique of nonnegative matrix factorization (NMF) [3]. On the other hand, in unsupervised SE methods, the learning of the application situations does not rely on the a priori noise traits. The well-known unsupervised methods include spectral subtraction (SS) [4], Wiener filtering [5] and short-time spectral amplitude estimation based on minimum mean-squared error criteria (MMSE-STSA) [6], just to name a few.

The new method proposed in this study applies the technique of nonnegative matrix factorization (NMF) to preserve the noise characteristics from the training set, and thus belongs to the supervised SE techniques. NMF is a well-known factor analysis approach that primarily deals with nonnegative data, and some SE methods [7] have used NMF to analyze the *intra-frame* magnitude spectra of speech signals. In contrast, in this study we propose using NMF to process the *inter-frame* temporal modulation spectra (temporal modulations in short) of speech signals to highlight the clean-speech portion and thus to realize speech enhancement, and the overall process is carried out on any single acoustic frequency or a set of neighboring acoustic frequencies. First, the (magnitude) temporal modulations of the spectrograms for either of clean speech and noise in the training set are respectively collected to form a data matrix. Next, the data matrix is factorized with NMF and accordingly we obtain the basis matrices corresponding to speech and noise, respectively. Finally, the temporal modulation of the testing noise-corrupted speech is NMF-encoded via a fixed basis matrix, which is the concatenation of clean speech and noise basis matrices. The partition corresponding to the clean-speech basis is extracted from the temporal modulation, and is then converted to the spectrogram and time-domain signal in sequence. The reconstructed time-domain signal is shown to have less noise in comparison with the original counterpart, which indicates the capability of the proposed method in speech enhancement.

II. PROPOSED METHOD

The time-varying dynamic of speech signals plays an important role in speech recognition and understanding, which can be characterized by the temporal modulation spectrum among others. In general, the temporal modulation spectrum is often referred to as the discrete Fourier transform (DFT) for the time series of short-time acoustic spectra. In this study, we present a novel NMF-based framework which enhances the magnitude part of the temporal modulation spectrum for a speech signal so as to reduce the noise effect.

The procedure of the presented SE framework is as follows. In the training stage, the magnitude parts of the temporal modulation spectra (with respect to any single acoustic frequency or a set of neighboring acoustic frequencies) for either of noise-free speech utterances and speech-free noise segments are collected and arranged as the columns of a data matrix $\mathbf{V}_u$, $u = \mathrm{s}, n$, representing the speech and noise respectively. Then the matrix $\mathbf{V}_u$ is analyzed by NMF, viz. $\mathbf{V}_u \approx \mathbf{W}_u\mathbf{H}_u$, in which $\mathbf{W}_u$ and $\mathbf{H}_u$ respectively denote the basis matrix and encoding matrix. As for the enhancement stage, the magnitude modulation spectrum of the of the tested noise-corrupted utterance, denoted by a vector $\mathbf{v}_y$, is analyzed via NMF, $\mathbf{v}_y \approx \mathbf{W}\mathbf{h}_y$, by keeping the basis matrix $\mathbf{W}$ fixed to be the concatenation of $\mathbf{W}_s$ and $\mathbf{W}_n$, i.e., $\mathbf{W} = [\mathbf{W}_s \quad \mathbf{W}_n]$. In other words, only the encoding vector $\mathbf{h}_y$ is iteratively updated in the NMF process. The NMF approximation for $\mathbf{v}_y$ can be written as follows,

$$\mathbf{v}_y \approx \mathbf{W}\mathbf{h}_y = [\mathbf{W}_s \quad \mathbf{W}_n]\begin{bmatrix}\mathbf{h}_s\\ \mathbf{h}_n\end{bmatrix} = \mathbf{W}_s\mathbf{h}_s + \mathbf{W}_n\mathbf{h}_n, \qquad (1)$$

where $\mathbf{h}_s$ and $\mathbf{h}_n$ denote the components in the encoding vector $\mathbf{h}_y$ associated with $\mathbf{W}_s$ and $\mathbf{W}_n$, respectively. Finally, the enhanced temporal modulation spectrum is expressed as a scaled version of $\mathbf{v}_y$:

$$\tilde{\mathbf{v}}_y = (\mathbf{W}_s\mathbf{h}_s./(\mathbf{W}_s\mathbf{h}_s + \mathbf{W}_n\mathbf{h}_n)).\times\, \mathbf{v}_y, \qquad (2)$$

where the symbols "./" and ".×" denote the element-wise division and multiplication operations, respectively.

III. Experimental Setup

The evaluation experiments are conducted on the utterances included in the Aurora-2 database [8], which contains English digit strings at 8 kHz sampling rate. As to the training, the clean-speech NMF-basis matrix $\mathbf{W}_s$ is obtained from 39 utterances produced by a single female speaker (labeled "FAK" in Aurora-2 dataset), and the noise matrix $\mathbf{W}_n$ is from the siren noise. Please note that both $\mathbf{W}_s$ and $\mathbf{W}_n$ are produced with respect to each acoustic frequency. In addition, the testing data consists of 10 utterances produced by the same speaker "FAK" as the training set, and each utterance is corrupted by siren noise at five signal-to-noise ratios (SNRs): 20 dB, 15 dB, 10 dB, 5 dB and 0 dB.

In particular, the level of enhancement for the processed utterances is evaluated via the metric of perceptual estimation of speech quality (PESQ) [9]. PESQ reflects the quality difference between the enhanced and clean speech signals, which ranges from 0.5 to 4.5. A higher PESQ score indicates that the enhanced utterance is closer to its clean counterpart.

IV. Experimental Results and Discussions

To begin with, our presented method is conducted in the manner that each acoustic-frequency temporal modulation spectrum is processed individually. The upper part of Table 1 shows the PESQ results for the signals enhanced via the presented method (the cases of "unprocessed" and "$N = 1$"). The results with respect to unprocessed signals are also listed in this table for comparison. From these results, we have the following two observations:

1. Noise causes a significant degradation to the speech signals and results in low speech quality, and it is of no surprise that the PESQ score decreases as the SNR worsens.
2. The speech signals enhanced by our method achieve higher PESQ scores relative to the unprocessed counterparts, indicating the effectiveness of the presented method. In particular, the presented method behaves better (giving higher PESQ improvement) when the noise level is higher, which reveals that the idea of factorizing for the temporal modulation spectrum helps to separate the noise component from the noisy speech.

As for the second part, we proceed the presented method on the neighboring acoustic-frequency components together. That is, the temporal modulation spectra of the neighboring acoustic frequencies share the same NMF-wise speech and noise basis matrices for noise separation. The corresponding PESQ results for different assignments for the number of neighboring acoustic frequencies jointly processed (the cases of $N = 2$ and $N = 4$) are listed in the last two rows of Table 1. Observing these results, we find that the speech quality can be also improved obviously when we process the neighboring acoustic frequencies jointly, and the degree of improvement is almost identical, sometimes better than, the single acoustic frequency processing. The probable explanation is: the neighboring acoustic frequency components are mutually correlated, and they give rise to more training data for preparing the NMF basis matrices of clean-speech and noise in comparison with the single acoustic frequency processing. Moreover, this promising result suggests that our method can be implemented at a relatively low acoustic frequency resolution without affecting its enhancement performance.

Table I

The PESQ Scores for the unprocessed Noisy utterances and the utterances enhanced by the Proposed method. The variable N denotes the number of acoustic frequency components being processed together

		0 dB	5 dB	10 dB	15 dB	20 dB
unprocessed		2.4004	2.4941	2.7811	2.9304	3.2094
Enhanced	$N = 1$	2.5339	2.6356	**2.8811**	3.1078	3.3130
	$N = 2$	**2.5360**	**2.6419**	2.8723	3.1124	**3.3191**
	$N = 4$	2.5156	2.6417	2.8629	**3.1127**	3.3086

V. Conclusions and Future Works

This study presents and evaluates a novel NMF-based speech enhancement method that processes the temporal modulation spectrum of speech signals. The preliminary evaluation reveals a promising result for this novel method. As for the future avenue, we will do more thorough evaluations for this method and extend it to deal with the mismatched speaker and noise situations.

References

[1] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), pp. 163–176, 2006.

[2] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), pp. 882–892, 2007.

[3] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proceedings of ICASSP*, 2012.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), pp. 113–120, 1979.

[5] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, 14(6), pp. 2098–2108, 2006.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, 32(6), pp. 1109–1121, 1984.

[7] Hao-teng Fan, Jeih-weih Hung, Xugang Lu, Syu-Siang Wang and Yu Tsao, "Speech enhancement using segmental nonnegative matrix factorization", in *Proceedings of ICASSP*, 2014

[8] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millennium*, pp. 181-188, 2000.

[9] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, 2001.